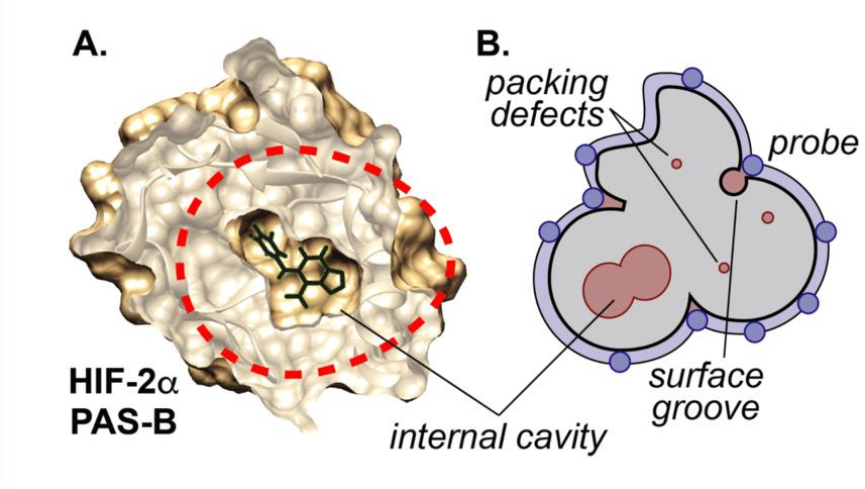
Cavities in Protein Structures



Huimin Lu
BINP 39 (30 Credits)

Background

- **Protein cavities are specific regions on the protein core**, and they are highly related to ligand binding; molecular transport; and enzyme catalysis.
- **Structure-based drug design** is a potential application for cavities.
- Variants in protein cavities **were not be widely investigated before**.
- In this project we investigated **variants in protein cavities**

The scientific problem addressed

This project use variant dataset from *VariBench* which includes **amino acids substitutions** in disease-related proteins.

Specific task:

- **Identifying cavities** in proteins using *CICLOP*
- **Locating variants** in disease-related proteins
- Analysis of **cavities** and **variants** located in protein cavities

Importance of the problem

This project investigated whether the **amino acid distribution** between **protein cavities** and **variants in protein cavities** are significantly different.

State of the art of existing solutions

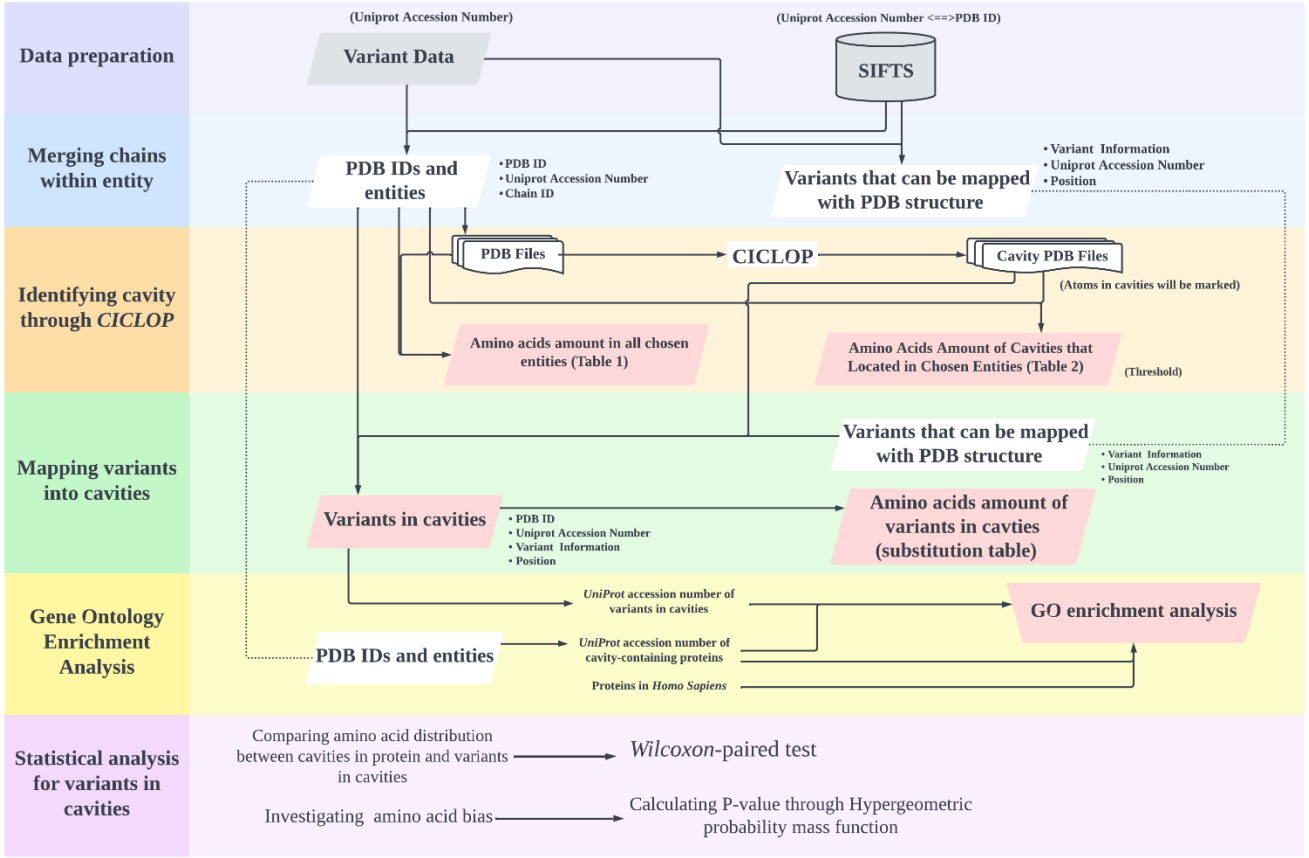
The **amino acids** and **variants** in protein cavities haven't been studied before.

The tools used to address the problem and the motivation for that choice

- *CICLOP*: a tool to locate atoms in cavities. There are several tools to identify cavities in proteins, but *CICLOP* was chosen because of its **high throughput capability**.
- *SIFTS*: a database that maps *UniProt* and PDB entries at residue-level. It also has API.
- *PDBSWS*: an API that offers mapping information between PDB chains and *UniProt*.
- Proprietary programs: were needed at many stages of project.

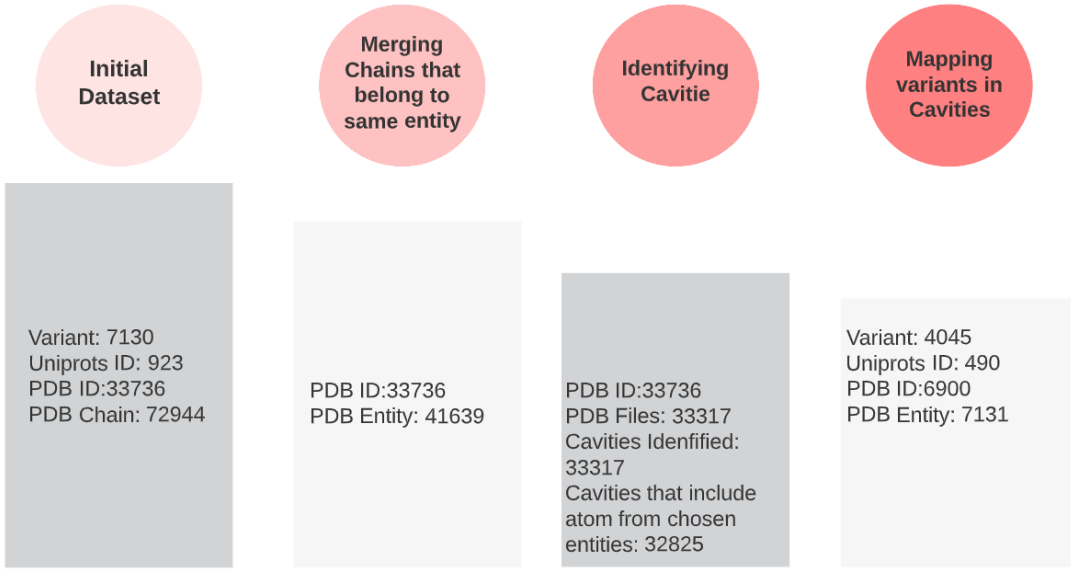
Result

- Constructed **workflow** for discovery of **cavities** and **variants** in protein cavities



Cavity Identification

- 32 825 **cavities containing proteins** identified in 33 736 PDB IDs.



Entity ID: 2	
Molecule	**Chains**
BLOOD COAGULATION FACTOR VIIA heavy chain	B [auth H]

B is shown in 'mmCIF' file
H is shown in PDB file

Entity ID: 1	
Molecule	**Chains**
Plasminogen activator inhibitor 1	A, B

A, B are identical.

- Gene Ontology Enrichment Analysis for cavity proteins in three branches of the ontology

Molecular Function for cavity-containing proteins

GO_Enrichment_Analysis_in_Variant.R | draft.Rmd | Statistics_Analysis.Rmd | GO_Enrichment_Analysis_in_Variant.Rmd | Cavity_MF

Filter

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0033218	GO:0033218	amide binding	157/3916	400/18368	1.414394e-16	1.707173e-13	9.781649e-14	7099/322/773/32/9854/9536/1200/6892/8811/3061/7097/1...	157
GO:0004713	GO:0004713	protein tyrosine kinase activity	71/3916	136/18368	1.943116e-15	1.172670e-12	6.719089e-13	2051/8767/8444/3717/8576/25/1956/2064/4914/3643/3932...	71
GO:0019199	GO:0019199	transmembrane receptor protein kinase activity	73/3916	143/18368	3.704927e-15	1.490615e-12	8.540831e-13	2051/8767/8444/3717/8576/25/1956/2064/4914/3643/3932...	73
GO:0004674	GO:0004674	protein serine/threonine kinase activity	148/3916	386/18368	1.063372e-14	3.208726e-12	1.838515e-12	6446/29904/10733/3551/8573/2580/375449/3656/8767/84...	148
GO:1901681	GO:1901681	sulfur compound binding	111/3916	265/18368	2.470435e-14	5.963629e-12	3.417001e-12	32/9536/2660/2255/10563/9731/8435/80739/30008/2147/4...	111
GO:0042277	GO:0042277	peptide binding	126/3916	321/18368	1.433428e-13	2.883580e-11	1.652215e-11	7099/322/773/9854/9536/1200/6892/8811/3061/7097/1026...	126
GO:0004714	GO:0004714	transmembrane receptor protein tyrosine kinase activity	63/3916	124/18368	3.718939e-13	6.412513e-11	3.674200e-11	2051/8767/8444/3717/8576/25/1956/2064/4914/3643/3932...	63
GO:0051287	GO:0051287	NAD binding	37/3916	56/18368	6.815314e-13	1.028261e-10	5.891660e-11	26227/3420/7358/4720/3417/3939/216/2746/1727/217/127...	37
GO:0140272	GO:0140272	exogenous protein binding	45/3916	77/18368	1.443947e-12	1.936494e-10	1.109559e-10	4864/1956/3949/920/2993/213/7037/3690/3383/1604/3678...	45
GO:0001618	GO:0001618	virus receptor activity	44/3916	76/18368	4.006579e-12	4.720976e-10	2.704994e-10	4864/1956/3949/920/2993/7037/3690/3383/1604/3678/468...	44
GO:0005543	GO:0005543	phospholipid binding	163/3916	466/18368	4.302463e-12	4.720976e-10	2.704994e-10	9743/3092/7287/5286/6861/2494/889/6386/9854/1200/533...	163
GO:0002020	GO:0002020	protease binding	64/3916	135/18368	1.283006e-11	1.290491e-09	7.394168e-10	8797/8767/5071/8996/9507/8915/462/5265/1471/3949/127...	64
GO:0140030	GO:0140030	modification-dependent protein binding	72/3916	160/18368	1.565065e-11	1.453103e-09	8.325894e-10	8805/23378/23476/5253/9682/100137047/3329/7409/5336...	72
GO:0019842	GO:0019842	vitamin binding	68/3916	148/18368	1.789015e-11	1.542387e-09	8.837467e-10	32/8566/5264/10558/9517/8974/8029/8985/8879/2806/213...	68
GO:0045296	GO:0045296	cadherin binding	123/3916	332/18368	2.792490e-11	2.247024e-09	1.287485e-09	1192/10528/1654/11122/9973/1500/8615/2317/3417/9217/...	123

column 2: character

Cellular Component for cavity-containing proteins

GO_Enrichment_Analysis_in_Variant.R | draft.Rmd | Statistics_Analysis.Rmd | GO_Enrichment_Analysis_in_Variant.Rmd | Cavity_CC

Filter

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0005759	GO:0005759	mitochondrial matrix	204/3926	480/19550	9.159035e-30	6.722731e-27	4.010693e-27	8050/5138/5442/4706/7015/9692/587/4968/2309/10469/99...	204
GO:0031983	GO:0031983	vesicle lumen	148/3926	327/19550	3.236765e-25	1.187893e-22	7.086813e-23	5768/6386/1654/2219/8566/317/5709/81/9535/8993/3417/...	148
GO:0060205	GO:0060205	cytoplasmic vesicle lumen	147/3926	325/19550	5.095769e-25	1.246765e-22	7.438035e-23	5768/6386/1654/2219/8566/317/5709/81/9535/8993/3417/...	147
GO:0034774	GO:0034774	secretory granule lumen	145/3926	322/19550	1.849125e-24	3.393145e-22	2.024305e-22	5768/6386/1654/2219/8566/317/5709/81/9535/8993/3417/...	145
GO:0031252	GO:0031252	cell leading edge	168/3926	422/19550	4.530039e-21	6.650097e-19	3.967360e-19	4641/7099/322/1654/3636/9051/10253/23396/1500/23380/...	168
GO:0045121	GO:0045121	membrane raft	139/3926	335/19550	1.609500e-19	1.687676e-17	1.006845e-17	4641/5348/8797/28514/6386/1200/3551/8871/4864/7097/3...	139
GO:0098857	GO:0098857	membrane microdomain	139/3926	335/19550	1.609500e-19	1.687676e-17	1.006845e-17	4641/5348/8797/28514/6386/1200/3551/8871/4864/7097/3...	139
GO:0005775	GO:0005775	vacuolar lumen	86/3926	174/19550	3.754205e-18	3.444483e-16	2.054933e-16	6386/1200/8029/1515/1727/12/718/7276/2638/2512/383/2...	86
GO:0030055	GO:0030055	cell-substrate junction	158/3926	425/19550	1.293271e-16	1.054734e-14	6.292404e-15	129446/6386/8573/4659/2580/247/726/81/23396/4868/101...	158
GO:0005925	GO:0005925	focal adhesion	155/3926	418/19550	3.289105e-16	2.414203e-14	1.440282e-14	129446/6386/8573/4659/2580/247/726/81/23396/4868/101...	155
GO:0009897	GO:0009897	external side of plasma membrane	154/3926	421/19550	1.626905e-15	1.085589e-13	6.476482e-14	7099/11119/2219/4065/8029/1515/23308/10159/19/2936/2...	154
GO:0045177	GO:0045177	apical part of cell	157/3926	435/19550	3.341015e-15	2.043587e-13	1.219177e-13	5348/28514/9854/1856/8972/5205/8029/1515/9026/10159/...	157
GO:0030139	GO:0030139	endocytic vesicle	129/3926	336/19550	4.903626e-15	2.768663e-13	1.651748e-13	4641/7476/1856/9146/6892/10618/8029/7097/23380/9727/...	129
GO:0030016	GO:0030016	myofibril	98/3926	231/19550	6.670755e-15	3.497381e-13	2.086492e-13	129446/8789/845/4659/23336/8557/81/9997/11155/23363/...	98
GO:0043202	GO:0043202	lysosomal lumen	54/3926	97/19550	1.037748e-14	5.078047e-13	3.029496e-13	1200/8029/1515/2638/2717/3073/1509/5660/3074/1514/15...	54

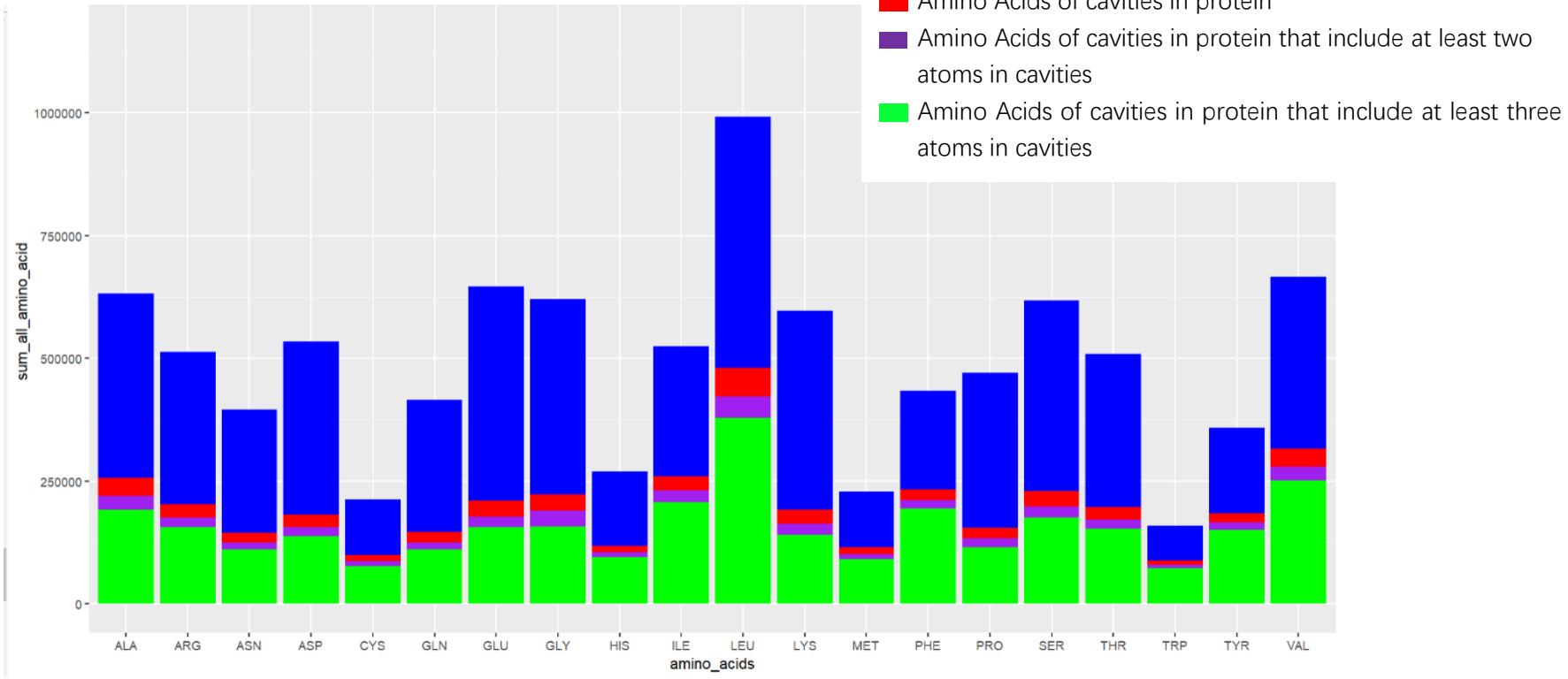
Biological Process for cavity-containing proteins

GO_Enrichment_Analysis_in_Variant.R | draft.Rmd | Statistics_Analysis.Rmd | GO_Enrichment_Analysis_in_Variant.Rmd | Variant_erich_go_BP | Cavity_BP | hyper_1

Filter

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0018108	GO:0018108	peptidyl-tyrosine phosphorylation	177/3871	375/18723	3.941144e-31	1.497624e-27	6.727336e-28	1856/2051/2255/8767/8444/9467/3717/8576/11116/25/195...	177
GO:0042060	GO:0042060	wound healing	192/3871	422/18723	4.739317e-31	1.497624e-27	6.727336e-28	7099/1192/7476/8573/2255/3717/8324/8291/80739/2157/2...	192
GO:0018212	GO:0018212	peptidyl-tyrosine modification	177/3871	378/18723	1.352789e-30	2.849875e-27	1.280165e-27	1856/2051/2255/8767/8444/9467/3717/8576/11116/25/195...	177
GO:0001819	GO:0001819	positive regulation of cytokine production	203/3871	467/18723	1.774962e-29	2.804440e-26	1.259756e-26	3965/7099/5293/11119/1654/2219/247/8767/7305/7097/37...	203
GO:0042110	GO:0042110	T cell activation	208/3871	487/18723	6.334890e-29	8.007301e-26	3.596884e-26	3965/10326/5293/11119/8943/8600/4092/2175/9092/8767/...	208
GO:0050878	GO:0050878	regulation of body fluid levels	174/3871	379/18723	8.229359e-29	8.668258e-26	3.893786e-26	7099/1192/8835/2255/3783/3717/80739/2157/2162/2806/1...	174
GO:0033674	GO:0033674	positive regulation of kinase activity	200/3871	467/18723	4.945575e-28	4.465148e-25	2.005747e-25	7099/8797/1654/1856/8795/8600/8312/2051/10253/904/37...	200
GO:0002697	GO:0002697	regulation of immune effector process	158/3871	339/18723	3.767826e-27	2.976583e-24	1.337083e-24	3965/7099/117157/259197/4092/8767/7305/4068/8993/87...	158
GO:0007599	GO:0007599	hemostasis	117/3871	222/18723	3.273820e-26	2.298949e-23	1.032689e-23	7099/1192/3717/80739/2157/2162/2147/2158/2159/5340/2...	117
GO:0007596	GO:0007596	blood coagulation	115/3871	217/18723	4.633826e-26	2.928578e-23	1.315519e-23	7099/1192/3717/80739/2157/2162/2147/2158/2159/5340/2...	115
GO:0050817	GO:0050817	coagulation	116/3871	222/18723	1.426775e-25	8.197469e-23	3.682307e-23	7099/1192/3717/80739/2157/2162/2147/2158/2159/5340/2...	116
GO:0070663	GO:0070663	regulation of leukocyte proliferation	122/3871	245/18723	2.170613e-24	1.143190e-21	5.135214e-22	3965/7099/23495/2255/8767/7305/23308/56253/90865/48...	122
GO:0050863	GO:0050863	regulation of T cell activation	149/3871	329/18723	4.299584e-24	2.090260e-21	9.389457e-22	3965/10326/8943/8600/4092/2175/9092/8767/23308/639/1...	149
GO:0070661	GO:0070661	leukocyte proliferation	145/3871	318/18723	7.737502e-24	3.492929e-21	1.569026e-21	3965/7099/11119/8600/23495/2255/8767/7305/23308/101...	145
GO:0022407	GO:0022407	regulation of cell-cell adhesion	185/3871	448/18723	9.133266e-24	3.848150e-21	1.728591e-21	3965/10326/8943/8600/4092/9092/10653/8767/10563/371...	185

- Animo acid distribution in cavities (bar-chart)



Variants in cavities

- Gene Ontology enrichment analysis for **variants** in cavity-containing proteins.

Molecular Function for variants in cavity-containing proteins

GO_Enrichment_Analysis_in_Variant.R | draft.Rmd | Statistics_Analysis.Rmd | GO_Enrichment_Analysis_in_Variant.Rmd | Variant_MF | Variant_CC | Variant_BP

Filter

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0002020	GO:0002020	protease binding	23/477	64/3916	6.624210e-07	0.0002353355	0.0001929199	8767/5071/8996/462/5265/1471/3949/1281/2335/5621/745...	23
GO:0051087	GO:0051087	chaperone binding	17/477	39/3916	8.465306e-07	0.0002353355	0.0001929199	332/5071/6647/5340/3106/2244/2335/213/5621/7450/7157...	17
GO:0004175	GO:0004175	endopeptidase activity	36/477	132/3916	1.343886e-06	0.0002490668	0.0002041763	10753/27429/2147/1675/5340/2161/5972/4057/2160/354/1...	36
GO:0008233	GO:0008233	peptidase activity	45/477	185/3916	2.103543e-06	0.0002923925	0.0002396932	10753/27429/2147/1675/5340/2161/5972/7037/4057/2160/...	45
GO:0005102	GO:0005102	signaling receptor binding	91/477	483/3916	3.993531e-06	0.0004440807	0.0003640419	7099/11119/2660/3551/8573/5364/8767/5071/9046/1500/8...	91
GO:0030234	GO:0030234	enzyme regulator activity	73/477	370/3916	8.047679e-06	0.0007457516	0.0006113412	10059/1654/317/332/5364/8996/4040/8412/216/1956/462/...	73
GO:0046914	GO:0046914	transition metal ion binding	73/477	372/3916	9.858442e-06	0.0007830420	0.0006419106	4194/5071/8856/125/5053/6647/2157/760/920/213/7018/4...	73
GO:0061134	GO:0061134	peptidase regulator activity	24/477	79/3916	1.138323e-05	0.0007911347	0.0006485447	317/332/8996/462/5265/718/727/1471/2335/4057/5621/35...	24
GO:0017171	GO:0017171	serine hydrolase activity	22/477	70/3916	1.453653e-05	0.0008980347	0.0007361776	27429/2147/1675/5340/2161/4057/2160/354/1991/2155/43...	22
GO:0044877	GO:0044877	protein-containing complex binding	88/477	480/3916	1.888747e-05	0.0010501433	0.0008608709	10059/1654/5071/1956/4893/3265/3845/3949/920/3117/31...	88
GO:0042277	GO:0042277	peptide binding	32/477	126/3916	2.643709e-05	0.0013362748	0.0010954316	7099/1471/3949/3106/3117/3123/3119/1410/335/5621/310...	32
GO:0016491	GO:0016491	oxidoreductase activity	53/477	255/3916	3.725823e-05	0.0015574558	0.0012767481	23028/3417/9131/125/3939/216/2936/5053/6647/2157/217...	53
GO:0033218	GO:0033218	amide binding	37/477	157/3916	3.766200e-05	0.0015574558	0.0012767481	7099/32/1471/3949/3106/3117/3123/3119/1410/335/5621/...	37
GO:0008236	GO:0008236	serine-type peptidase activity	21/477	69/3916	3.921651e-05	0.0015574558	0.0012767481	27429/2147/1675/5340/2161/4057/2160/354/1991/2155/43...	21
GO:0016903	GO:0016903	oxidoreductase activity, acting on the aldehyde or oxo grou...	11/477	24/3916	4.406919e-05	0.0016334982	0.0013390850	216/217/1738/593/231/594/8659/218/8644/501/316	11

Cellular Component for variants in cavity-containing proteins

GO_Enrichment_Analysis_in_Variant.R | draft.Rmd | Statistics_Analysis.Rmd | GO_Enrichment_Analysis_in_Variant.Rmd | Variant_CC

Filter

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0030141	GO:0030141	secretory granule	76/477	311/3926	2.691726e-10	1.092841e-07	7.876841e-08	1654/317/3417/6647/2157/2162/4860/1675/5340/5265/718...	76
GO:0099503	GO:0099503	secretory vesicle	80/477	353/3926	4.199622e-09	5.920551e-07	4.267340e-07	10059/1654/317/3417/6647/2157/2162/4860/1675/5340/52...	80
GO:0031983	GO:0031983	vesicle lumen	44/477	148/3926	4.374791e-09	5.920551e-07	4.267340e-07	1654/317/3417/2157/2162/4860/1956/1675/5340/5265/718...	44
GO:0034774	GO:0034774	secretory granule lumen	43/477	145/3926	7.227006e-09	7.335411e-07	5.287125e-07	1654/317/3417/2157/2162/4860/1675/5340/5265/718/7040...	43
GO:0009986	GO:0009986	cell surface	76/477	334/3926	9.141408e-09	7.422823e-07	5.350129e-07	7099/11119/4040/2936/1956/2147/5340/718/3949/7040/92...	76
GO:0060205	GO:0060205	cytoplasmic vesicle lumen	43/477	147/3926	1.144184e-08	7.742312e-07	5.580406e-07	1654/317/3417/2157/2162/4860/1675/5340/5265/718/7040...	43
GO:1904813	GO:1904813	ficolin-1-rich granule lumen	21/477	53/3926	2.927429e-07	1.697909e-05	1.223797e-05	1654/317/3417/4860/1675/5265/1471/2934/1508/4318/272...	21
GO:0005783	GO:0005783	endoplasmic reticulum	95/477	495/3926	8.531733e-07	4.329854e-05	3.120818e-05	6820/4221/10059/27429/5071/8996/4040/2157/1956/2147/...	95
GO:0043235	GO:0043235	receptor complex	45/477	182/3926	1.204558e-06	5.433894e-05	3.916574e-05	7099/3551/5364/27429/1956/3949/2688/920/5284/7037/70...	45
GO:0101002	GO:0101002	ficolin-1-rich granule	25/477	79/3926	3.127636e-06	1.269820e-04	9.152451e-05	1654/317/3417/4860/1675/5265/1471/3689/2934/1508/160...	25
GO:0098552	GO:0098552	side of membrane	51/477	228/3926	5.699959e-06	2.103803e-04	1.516353e-04	7099/11119/3551/27429/2936/2147/5340/3845/3949/920/3...	51
GO:0030134	GO:0030134	COPII-coated ER to Golgi transport vesicle	14/477	32/3926	7.443084e-06	2.518243e-04	1.815068e-04	2157/5265/3106/3117/3123/3119/3105/3115/351/3133/311...	14
GO:0005739	GO:0005739	mitochondrion	89/477	479/3926	8.583356e-06	2.680648e-04	1.932124e-04	10059/32/8473/587/4968/27429/5071/8996/3417/2744/913...	89
GO:0072562	GO:0072562	blood microparticle	20/477	59/3926	9.680752e-06	2.807418e-04	2.023495e-04	2162/2147/5340/462/718/7040/335/2244/6521/325/2335/2...	20
GO:0098576	GO:0098576	lumenal side of membrane	9/477	15/3926	1.367371e-05	3.401434e-04	2.451643e-04	3106/3117/3123/3119/3105/3115/5476/3133/3113	9

Biological Process for variants in cavity-containing proteins

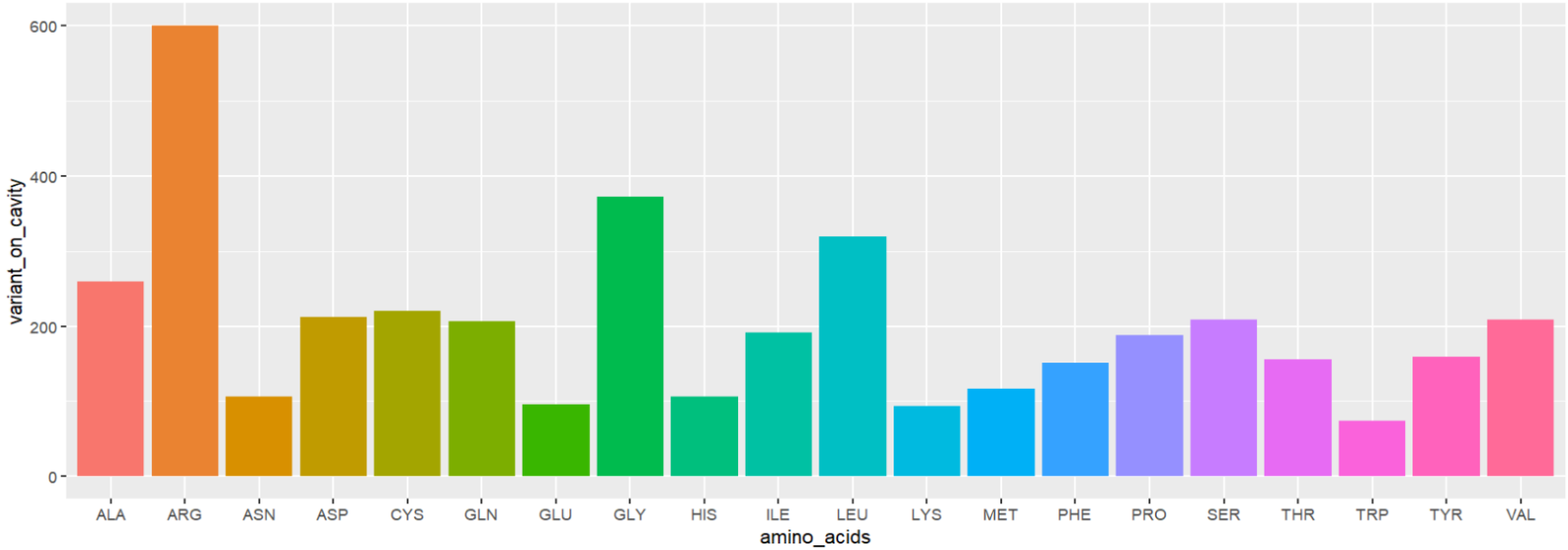
GO_Enrichment_Analysis_in_Variant.R | draft.Rmd | Statistics_Analysis.Rmd | GO_Enrichment_Analysis_in_Variant.Rmd | Variant_BP

Filter

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:1901698	GO:1901698	response to nitrogen compound	96/476	413/3871	3.265537e-11	4.980753e-08	3.384846e-08	7099/4221/2660/8473/4968/8767/5071/23028/4040/8412/9...	96
GO:0043069	GO:0043069	negative regulation of programmed cell death	81/476	325/3871	3.827310e-11	4.980753e-08	3.384846e-08	1654/4194/332/4968/6899/27429/5071/23028/8996/4040/1...	81
GO:0010243	GO:0010243	response to organonitrogen compound	91/476	384/3871	3.885143e-11	4.980753e-08	3.384846e-08	7099/4221/2660/8473/4968/8767/5071/23028/4040/8412/9...	91
GO:1901701	GO:1901701	cellular response to oxygen-containing compound	99/476	438/3871	8.607583e-11	8.276191e-08	5.624376e-08	7099/4221/10059/2660/8473/8767/6899/27429/5071/2302...	99
GO:0052548	GO:0052548	regulation of endopeptidase activity	49/476	157/3871	1.212397e-10	9.325758e-08	6.337646e-08	1654/317/332/8767/27429/8996/9131/462/5265/718/727/1...	49
GO:0043066	GO:0043066	negative regulation of apoptotic process	78/476	317/3871	1.866668e-10	1.073088e-07	7.292545e-08	1654/4194/332/4968/6899/27429/5071/23028/8996/4040/1...	78
GO:0001775	GO:0001775	cell activation	94/476	413/3871	1.953098e-10	1.073088e-07	7.292545e-08	7099/5293/11119/8767/6647/4860/3251/1956/2147/3949/7...	94
GO:0052547	GO:0052547	regulation of peptidase activity	50/476	165/3871	2.527780e-10	1.215230e-07	8.258523e-08	1654/317/332/8767/27429/8996/9131/462/5265/718/727/1...	50
GO:0009410	GO:0009410	response to xenobiotic stimulus	54/476	188/3871	4.310421e-10	1.841987e-07	1.251787e-07	32/4968/8856/8412/11200/3939/6647/4860/1956/5972/335...	54
GO:0010035	GO:0010035	response to inorganic substance	62/476	233/3871	6.295561e-10	2.241076e-07	1.523002e-07	4968/5071/9131/3939/6647/1956/760/1410/2244/7018/562...	62
GO:0042325	GO:0042325	regulation of phosphorylation	103/476	478/3871	6.488266e-10	2.241076e-07	1.523002e-07	7099/4221/1654/2660/3551/8473/8767/6899/5071/4040/84...	103
GO:0014070	GO:0014070	response to organic cyclic compound	85/476	367/3871	6.992437e-10	2.241076e-07	1.523002e-07	10059/32/2660/4968/8767/5071/23028/4040/3417/9131/11...	85
GO:0009611	GO:0009611	response to wounding	62/476	234/3871	7.614458e-10	2.252708e-07	1.530907e-07	7099/8573/8996/6647/2157/2162/1956/2147/5340/2161/46...	62
GO:0002460	GO:0002460	adaptive immune response based on somatic recombinatio...	44/476	140/3871	8.989771e-10	2.469619e-07	1.678316e-07	7099/8767/3251/718/727/3265/7040/920/3106/3117/3123/...	44
GO:1902533	GO:1902533	positive regulation of intracellular signal transduction	79/476	334/3871	1.096815e-09	2.812233e-07	1.911151e-07	7099/5293/10059/1654/3551/5364/8767/6899/5071/8412/6...	79

Two groups of Gene Ontology enrichment analyses showed that the annotations for **cavity-containing proteins** do not significantly differ from proteins that have variants in cavities.

- Amino acid distribution of variants



- Biased distribution of variants (hypergeometric probability mass function)

	sum_cavity_amino_acid_1	variant_on_cavity	P_value	Expected_Value	Expect_Text	Fold_Change_Value
A	257211	259	0.0256	258.142674885204	over-enriched	1.00332112896551
R	202387	600	9.26e-122	203.12009028382	over-enriched	2.95391755272272
N	144450	106	9.89e-05	144.97322971089	under-enriched	1.36767197840462
D	181046	212	0.00227	181.701788482089	over-enriched	1.16674690860788
C	99651	220	2.5e-26	100.01195786722	over-enriched	2.19973695837534
Q	146006	96	1.43e-06	146.534865885553	under-enriched	1.52640485297451
E	209596	207	0.0277	210.355202869392	under-enriched	1.01620870951397
G	223115	372	2.27e-21	223.923171664556	over-enriched	1.66128407897539
H	117724	106	0.0202	118.15042225327	under-enriched	1.11462662503085
I	260069	192	6.72e-07	261.011027190595	under-enriched	1.35943243328435
L	480260	319	1.42e-17	481.999607483226	under-enriched	1.51097055637375
K	191702	93	1.79e-16	192.396386860762	under-enriched	2.06877835334152
M	114974	117	0.037	115.390461147662	over-enriched	1.01394863003691
F	232058	151	9.6e-10	232.898565179991	under-enriched	1.542374603841
P	154774	188	0.00104	155.334625512449	over-enriched	1.21029036108201
S	229694	209	0.00944	230.526002251389	under-enriched	1.1029952260832
T	197565	156	0.000198	198.280623937916	under-enriched	1.27102964062767
W	87899	74	0.0135	88.2173895351852	under-enriched	1.19212688561061
Y	183799	159	0.00468	184.464760454357	under-enriched	1.16015572612803
V	316421	209	3.86e-12	317.567146544475	under-enriched	1.51946003131328

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{k}}$$

N: Sum of amino acids in cavities
K: The amount of one type of amino acid in cavities
n: Sum of amino acids in variant
k: The amount of one type of amino acid in variants

$$Expected\ Amount = \frac{n * K}{N}$$

$k = Expected\ Amount\ (Matched)$:

$$Fold\ Change = 1$$

$k < Expected\ Amount\ (Under - Representation)$:

$$Fold\ Change = \frac{Expected\ Amount}{k}$$

$k > Expected\ Amount\ (Over - Representation)$:

$$Fold\ Change = \frac{k}{Expected\ Amount}$$

- Substitution table

  Distribution of variants is highly biased, *arginine* is by far **the most frequently altered amino acid**. This is line with the high mutability of CpG dinucleotides that are common in codons for *arginine* and known as **mutational hotspots**.

Original Amino Acid
Substitution Amino Acid

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	sum_col
A	0	0	0	28	1	0	13	13	0	1	0	0	0	1	36	19	76	0	0	71	259
R	2	0	0	0	118	104	0	61	93	5	28	17	4	0	44	17	11	96	0	0	600
N	0	0	0	12	0	0	0	0	11	7	0	28	0	0	0	31	9	0	7	1	106
D	10	0	58	0	0	0	18	45	27	0	0	0	0	0	0	0	0	0	25	29	212
C	1	49	0	0	0	0	0	17	0	0	0	0	0	31	0	29	0	11	82	0	220
Q	0	32	0	0	0	0	9	0	16	0	5	14	0	0	20	0	0	0	0	0	96
E	15	0	0	23	0	19	0	37	0	0	0	107	0	0	0	0	0	0	0	6	207
G	27	99	1	53	17	0	52	0	0	0	0	0	0	0	0	64	0	4	0	55	372
H	1	36	4	8	0	11	0	0	0	0	8	0	0	0	13	0	0	0	25	0	106
I	0	1	20	0	0	0	0	0	0	0	7	2	17	20	0	15	70	0	0	40	192
L	4	39	0	0	0	18	0	0	10	4	0	0	9	47	115	25	0	6	0	42	319
K	1	25	16	0	0	8	28	0	0	3	0	0	3	0	0	0	9	0	0	0	93
M	0	12	0	0	0	0	0	0	0	20	9	15	0	0	0	0	29	0	0	32	117
F	2	0	0	0	22	0	0	0	0	11	56	0	0	0	0	34	0	0	7	19	151
P	11	19	0	0	0	4	0	0	10	0	79	0	0	0	0	45	20	0	0	0	188
S	6	21	17	0	15	0	0	13	1	10	35	0	0	25	40	0	7	5	14	0	209
T	30	8	7	0	0	0	1	0	0	47	0	4	29	0	17	13	0	0	0	0	156
W	0	25	0	0	25	0	0	9	0	0	6	0	0	0	0	9	0	0	0	0	74
Y	2	0	13	14	69	0	0	1	33	0	0	0	0	9	0	16	0	2	0	0	159
V	31	0	0	10	0	1	12	17	0	40	24	0	52	22	0	0	0	0	0	0	209
sum_row	143	366	136	148	267	165	133	213	201	148	257	187	114	155	285	317	231	124	160	295	4045

Implications

- The bias in amino acid distribution could be used in **variation interpretation.**
- The data can be applied in studies of **structural bases and mechanisms of diseases**.
- A further application could be **protein engineering**
- These data can detect potential target sites for **drug design**

Future steps

- Discover **other features** for variants locating in cavities, for example: the **volume** of cavities; **secondary structural elements, etc.**
- Simplifying the steps and packing all steps into one workflow using e.g. **snakemake**, making it work automatically.